Testing Transferability of Defense-GAN to Physical Adverserial Examples

Rohan Pradhan - Bibin Sebastian - Karthik Raja - Chamanpreet Kaur - Mihir Pathare

The research area to combat and defend against adversarial examples is gaining tremendous attention and generative adversarial networks (GANs) are proving to be a a potential breakthrough for presenting a solution to this problem. Therefore, we plan to explore and test using GANs as viable defense mechanism and furthermore test its transferability of learning to contend against a physical attack scenario.

The majority of research in attacking machine learning models is centered around obfuscating image classification techniques. Many of the attack scenarios and defence mechanisms assume adversarial access to the raw pixel data. However, in practical scenarios physical world image alterations are equally significant; therefore, they need to be detected and accounted for. As motivation, we could think about self driving cars using vision based systems to detect traffic lights and other standard information from road sign boards. This perceived information is often used for decision making and the cost of wrong decisions is very high and unacceptable. In these scenarios, it is often difficult to access the pixel values of the image, while simpler to make modifications to the physical world, like a road sign. Therefore, we are interested in extending the research work carried out by Samangouei et al.,2018; on utilizing a GAN as a filter on image classification problems. We intend to research and understand how much of the digital adversarial examples get transferred through the physical transformation and verify if state-of-the-art defense mechanisms, like Defense-GAN, can still generalize well against physical adversarial examples.

Related Works After performing a comprehensive literature review, we consider the following papers to serve as a baseline for our project goals:

- Samangouei, Pouya, Maya Kabkab, and Rama Chellappa. "Defense-GAN: Protecting classifiers against adversarial attacks using generative models." arXiv preprint arXiv:1805.06605 (2018).
- Kurakin, Alexey, Ian Goodfellow, and Samy Bengio. "Adversarial examples in the physical world." arXiv preprint arXiv:1607.02533 (2016).

We intend to first recreate the results of both these papers, then combine the work of both these papers to extend prior research.

Experimental Setup As a control experiment, we intend to recreate the results of both the papers. As per the attacks described in Kurkin et al paper, we intend to attack through a digital medium using FGSM as the baseline, iterative, and least likely targeted attacks. Furthermore, we will recreate the results of the Defense-GAN paper, through a digital medium, using the three attacks described above. Once the results of both reference papers have been recreated, we intend to extend these works by combining aspects of both these papers. We will test the effectiveness of the Defense-GAN mechanism under attack of adversarial examples which has undergone transformation through the physical medium. See Appendix for the experiments we plan to carry.

Appendix

Proposed Benchmark Experiments (Reproducing Previous Papers)

Kurakin	Digital Clean Example on model M
	Physical Clean Example on model M
	Digital Adversarial Example using Attack A_1 on model M
	Digital Adversarial Example using Attack A_2 on model M
	Digital Adversarial Example using Attack A_3 on model M
	Physical Adversarial Example using Attack A_1 on model M
	Physical Adversarial Example using Attack A_2 on model M
	Physical Adversarial Example using Attack A_3 on model M
Samangouei	Clean Example on model M with Defense-GAN
	Digital Adversarial Example using Attack A_1 on model M with Defense-GAN
	Digital Adversarial Example using Attack A_2 on model M with Defense-GAN
	Digital Adversarial Example using Attack A_3 on model M with Defense-GAN

Extensions of work (New experiments to be run)

Extensions	Physical Clean Example on model M with Defense-GAN
	Physical Adversarial Example using Attack A_1 on model M with Defense-GAN
	Physical Adversarial Example using Attack A_2 on model M with Defense-GAN
	Physical Adversarial Example using Attack A_3 on model M with Defense-GAN

Compiled on October 21, 2018